DOCUMENT RESUME

ED 291 806 TM 011 124

AUTHOR Green, Donald Ross
TITLE Sex Differences in Item Performance on a Standardized Achievement Battery.
PUB DATE 28 Aug 87
NOTE 24p.; Paper presented at the Annual Meeting of the American Psychological Association (95th, New York, NY, August 28-September 1, 1987). Some tables contain light/broken type.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; Educational Testing; Elementary Secondary Education; Item Analysis; Performance Factors; *Sex Differences; *Standardized Tests; *Test Bias; Testing Programs
IDENTIFIERS LOGIST Computer Program

ABSTRACT

Differential functioning of males and females on achievement test items was studied in a sample of 110,000 students in kindergarten through grade 12. First, item parameters and ability estimates for these parameters were obtained from LOGIST computer program runs for the entire group. Predicted performance for each item for each group was separately determined, and expected performance was then subtracted from observed performance for the item. Items with extreme values were examined for content bias. Females had higher mean scale scores on all but three of 72 tests in the basic battery. Differences favoring females increased up to grade 8 and then diminished in high school. Variations of scale scores for females were smaller than those for males. The number of items biased in favor of females was less than the number biased in favor of males. Until such time as these results are better explained, test constructors are advised to continue carefully balancing the number of male and female names and actors in passages and items. Test data are summarized in attached tables. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED291806

Sex Differences in Item Performance on a Standardized Achievement Battery

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

by

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
DONALD ROSS GREEN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Donald Ross Green

CTB/McGraw-Hill

August 28, 1987

Paper presented at the Annual Meeting of the
American Psychological Association
New York City

TM 011 124

BEST COPY AVAILABLE

Purpose of the Study

Differences between groups on item performance can be examined in two ways. The most common approach is to compare group performances on individual items among examinees obtaining comparable scores. Differences found are thus attributable to differential interactions between the processes elicited by the item and the functioning of the group apart from overall ability on the trait being measured. However, overall group differences in ability are obviously relevant.

In this study both approaches were used to assess differential functioning on achievement test items of males and females in a large (110,000) representative sample of students in grades K-12. First item parameters and ability estimates (thetas) for the three parameter model were obtained from LOGIST (Wingersky, Barton & Lord, 1982) runs using the entire group. Then predicted performance (E_i on each item for each group separately was determined using these ability estimates and these parameters both based on the total group. This expected performance was then subtracted from that observed for that item (i.e., O_i-E_i). The specific procedure followed was that suggested by Linn and Harnish (1981). The items were then ranked by the size of the female - male differences between these O-E values, giving $D_i = (O_{fi} - E_{fi}) - (O_{mi} - E_{mi})$.

Strong positive values of D indicate differential item functioning which favors females since this will occur only when females do better than expected, i.e., $(O_f - E_f) > 0$ and the male O-E value is either relatively small or negative. Similarly, negative values of D indicate differential functioning in favor of males. Note that of necessity overall these two values balance, given the procedure and relatively equal numbers of males and females, that is, the sum of the Ds for all items should be about zero. Nevertheless many individual items may exhibit large O-E values for one group, but not for the other.

Items which function in this manner have customarily been called biased. As has been noted many times (e.g., Green, 1975; Shepard, 1982), this does not necessarily mean unfairness, but in some circumstances it means just that. For convenience then the term bias will be used to indicate strong differential item functioning but it is not intended to indicate unfairness which must be judged independently.

To compare overall group differences, the female minus male p-values were obtained, i.e., $p_i = p_{if} - p_{im}$. Thus every item has two ratings, D_i and p_i. For each test the items with extreme values on these two indices are examined and categorized by content, format, and the presence or absence of male or female names, illustrations, and stereotypic activities. The intent of the content analyses is to see if the bias data shed any light on the uniformly higher scores of the female groups and vice versa. As just noted, the bias procedure is conditional on overall score in any particular test so that the two sorts of data are not necessarily related. Thus any content connections found may be enlightening.

Test Scores

The data come from the Spring 1985 standardization of the California Achievement Tests, Form E (CTB/McGraw-Hill, 1985). The sample consists of roughly 110,000 cases with valid scores drawn from a representative sample of schools around the United States. The comparisons reported are between males and females in the total sample as well as separately within the ethnic groups which were teacher- or self-identified as Black, Hispanic or Other (i.e., non-Black, non-Hispanic and non-Asian).[1] The number of students in these various groups are shown in Table 1. The mean scale scores and standard deviations are shown in Table 2.

Two general and presumably relevant characteristics of the male and female scores on this battery may be noted. First females have higher mean scale scores on all but three of the 72 tests in the basic battery.[2] The exceptions are Reading Vocabulary at kindergarten (Level 10) and at grade 6 (Level 16), and Mathematics Concepts and Applications at grades 11 and 12 (Level 20), but those differences are small. Table 3 shows the effect sizes for the mean differences. Effect size is the standardized difference which uses the average of the female and male variances (following Stanley, 1987).

Generally speaking it can be seen that the differences favoring females increase from kindergarten (Level 10) to grade 8 (Level 18) and then diminish in high school (Levels 19 and 20). The differences are much larger in some areas than others. Relatively high are Spelling, Language Mechanics, Language Expression and Mathematics Computation. Low are Reading Vocabulary and Mathematics Concepts and Applications. The male/female effect sizes within the three ethnic groups considered separately are similar. For comparison it may noted that effect sizes summarizing ethnic group differences in performance are about twice as large as these sex difference effect sizes (CTB/McGraw-Hill, 1987).

It may also be relevant to note that thirty years ago it was reported that on the 1957 Edition of the California Achievement Tests little difference in performance by sex could be found except in Language Mechanics where females outscored males in three of the four levels examined (Clark, 1958).

The second general characteristic is that the variances of the scale scores for the female groups are smaller than those for the male groups in 68 of the 72 cases. Table 4 shows the variance ratios. The ratios tend to increase with grade. As with effect sizes, Reading Vocabulary exhibits the lowest variance ratios, but for this statistic, unlike effect size, Mathematics Concepts and Applications stand as high as the rest of the tests.

ITEM DATA

Given the mean scale score differences in performance, it is not surprising that most of the items show differences in performance favoring the females. It is evident that no small subset of items in any of the tests create these overall differences. Nevertheless, to keep matters simple, it seemed desirable to focus on those items with "large" differences in difficulties. Unlike the score data just discussed which are in scale score units, the item data are in raw score units. Large was chosen arbitrarily to be .05, mostly because it seemed hard to defend any difference smaller than that as important. Note that, given the size of the groups, all differences of .04 or more are clearly improbable by chance. Using the .05 difference in p-value criterion, 1233 of the 3102 items in these tests favor females; none were found to favor males. See Table 5 for a listing of the frequency of these items.

As noted earlier, the differences between expected and obtained percent correct for males and for females have to more or less balance out since the numbers of cases are pretty nearly the same for the two groups on all tests; there are slightly more males than females in the elementary grades and slightly more females than males at the high school level. Thus generally speaking the expectation is that the obtained minus expected differences will be zero and that the number of differences favoring either group overall - the Ds defined earlier - will be about equal. To avoid spending effort in interpreting chance and/or unimportant differences it seems reasonable to set a standard of what constitutes a large net difference. As was done for the p-value differences, the rather arbitrary value chosen was .05; at .07 the number of items is very small and at .03 chance becomes more than a remote problem. Thus, items exceeding a D value of .05 were flagged as favoring females; items with D values less than -.05 were flagged as favoring males.

Table 6 shows the number of items in each test that favored one group or the other to this degree. Overall, 298 of the 3102 items being considered functioned quite differently for the males than they did for the females and will be called biased as defined above. As was true for the other phenomenon, there were more biased items in the higher levels of the test and more in some content areas than others; here the tests showing the most effects were the Reading Vocabulary tests. In the large majority of instances the tests had more items biased in favor of males than in favor of females, although the discrepancies were typically quite small.

Table 7 shows the number of items in each test having extreme values for both of these indices (i.e. both over |.05|). It can be seen that 97 items have such extreme indices with the majority being biased in favor of males even though females outscored them on all of these items. Neither Reading Vocabulary nor Mathematics Concepts and Applications included any such items while Language Mechanics had the most.

Also of interest are the items previously categorized by the test development editors as measuring higher order thinking skills in the Reading Comprehension, Language Expression, and Mathematics Concepts tests (CTB/McGraw-Hill, 1986). The frequencies of these items can be found in Table 8. Such items constitute nearly 40% of the items in these tests from grade 3 on. Almost 40% of the items in these tests that were biased in either direction were in this group of items. Of the 461 items requiring higher order thinking skills, 215 were in the group of items on which females performed much better than males, i.e., the differences in p-values exceeded .05. This class of items does not appear to be related to either bias or the differences in general level of performance.

CHARACTERISTICS OF EXTREME ITEMS

The extreme items other than those in Spelling were examined in several ways. First it was determined whether or not they belonged to the group of items requiring higher order thinking skills, and then they were categorized by objective. Table 9 shows the results of these counts. Plainly most of them are not among the items judged to involve higher order thinking skills. Of the 15 items that did involve higher order thinking skills, 12 were biased in favor of males.

Among the categories of instructional objectives (see Table 9) the only distinctive results are the bias favoring the female group in proofreading skills and perhaps that favoring the male group in punctuation skills. The proofreading result may be confounded with letter writing content, which, as described below, favors females.

Each of the 66 extreme items in Reading Comprehension, Language Mechanics, and Language Expression, was examined by four people (in a few cases six) looking for clues that might explain their differential functioning. Spelling and Computation were judged to be neutral. As far as format is concerned the only possibility found was a set of four items requiring the student to fill in the blank, each of which appeared to be biased in favor of males. However, there were a number of other such items that did not appear biased either way.

Agreement among the four judges was reached on a number of items categorized with respect to their context or content and as to the sex of the principal actors. The frequency of these items are shown in Table 10. Thirty-two of these extreme items could be categorized as to context and 32 as to sex of the principal actors or protagonists without any disagreements among the judges. Sixteen of these items appeared in both of these categorizations. Most of the remaining 18 items appeared neutral in all respects.

Most of the categories of content familiarity (context) used are traditional and stereotypic and the results appear to confirm the relevance of some of these stereotypes. Thus 24 of the 32 fell in

categories in which an expected direction could be asserted. Letter writing and literary imagery might be expected to favor females while items dealing with sports, or science and things mechanical, or construction might be expected to favor males; 23 of these 24 were biased in the expected direction. Examples were eight items requiring proofreading of letters, some to aunts written by loving nieces, all biased in favor of females and showing very large differences in p-values. Another set of items in Reading Comprehension concerned the history of football. The items asking about specific game actions referred to in the passage were both biased in favor of males; the other items in the set, not counted as biased, nevertheless tended in that direction. Note that the females still got these items right more often than males.

Other categories of context which might be relevant with more items to examine were history and geography. Although no substantial basis for any expectations are known for these categories, eight such items are also reported in Table 10.

Lastly there were 32 items in which the principal protagonist or actors were of one sex. Twenty-four of these were biased in the expected direction; that is, those items favoring females had female actors, those favoring males had male actors. In two of the other eight instances two judges felt that the results could be explained by a distractor which would draw counter to the otherwise expected direction of bias for the item.[3]

DISCUSSION

The noteworthy facts reported here seem to be:

1. Females score consistently higher than males on most of the tests. This is true in all ethnic groups examined.
2. Where this is the case the higher scores result from better performance on almost all the items in the test; no small group of items account for the phenomenon.
3. Male performance is more variable than that of females on a large majority of the tests. This is also true in all ethnic groups whose data were examined.
4. The number of items biased in favor of females is less than the number biased in favor of males. The total number is less than 10% of the items.
5. A substantial number of the items with both large p differences and large Ds are biased in the direction one would expect on the basis of differential content familiarity in the two groups or identification with sex of the principal persons identified in the material.

Nothing in the data presented seems to explain anything about the first four facts just listed. In particular it does not appear that the information about performance helped interpret the bias data although that cannot be determined until the 200 biased items not showing large p differences are examined.

The fifth set of facts may be considered understandable but it is a bit surprising that the results on context are so clear. Perhaps looking only at items with large Ds is responsible. It is also surprising to me that an item can be biased merely because of the sex of the person described in the item. Perhaps this is a type I error, although the probability of that was about .02; Johnson (1984) reported results and summarized other studies which would support the chance interpretation.

It may also be that other factors played a role. Given the results reported by Schmitt (1986) showing that language can provide a convincing explanation for differential item performance in some instances, one might speculate that somehow language differences between males and females of the sort described by Smith (1985) could be one such factor.

In any case these data do suggest that until good explanations are forthcoming test constructors would be well advised to continue the common policy of carefully balancing the number male and female names and actors in passages and items.

FOOTNOTES

[1]The number of cases in the Asian group was too small to allow separate analyses.

[2]To reduce the number of items and tests to be examined, thirty-three of the tests in the battery are not included in this report; They include the Word Analysis, Study Skills, Social Studies, and Science tests. The two latter sets of tests had very small numbers of cases and the data were judged not to be worth analysis. Nevertheless it may be noted that the nine Science tests had mean differences favoring males, while all the other omitted tests favored females.

[3]The author would like to thank Evonne Schaeffer, Sally Shay and Daniela Johnson for their assistance and good judgment in dealing with the classification of items.

REFERENCES

CTB/McGraw-Hill (1985). California achievement tests, form E. Monterey, CA: Author.

CTB/McGraw-Hill (1986). H.O.T.S. scores. Monterey, CA: Author.

CTB/McGraw-Hill (1987). California achievement tests, forms E and F, technical report. Monterey, CA: Author.

Clark, Willis W. (1958, March). Research findings on mental sex differences and their implications for education: An analysis of sex differences found in mental ability & achievement test results. Presentation at the American Association of School Administrators, Cleveland, Ohio.

Green, D. R. (1975). What does it mean to say a test is biased. Education & Urban Society, 8(1), pp. 33-52.

Johnson, E. S. (1984). Sex differences in problem solving. Journal of Educational Psychology, 76(6), pp. 1359-1371.

Linn, R. L. & Harnish, D. L. (1981). Interactions between item content and group membership on achievement test items. Journal of Educational Measurement, 18(2), pp. 109-118.

Schmitt, A. P. (1985). Assessing unexpected differential item performance of Hispanic candidates on SAT form 3FSA08 and TSWE form E47. ETS Statistical Report (SR-85-169). Princeton, NJ: Educational Testing Service.

Shepard, L. A. (1982). Definitions of bias. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 9-30). Baltimore, MD: Johns Hopkins.

Smith. P. M. (1985). Language, the sexes and society. Oxford: Basil Blackwell.

Stanley, Julian C. (1987, April). How "effect size" was computed for the studies reported in this symposium. In J. J. Michael (Chair), Sex differences in cognitive abilities and achievements. Symposium conducted at the American Educational Research Association, Washington, D.C.

Wingersky, M. S., Barton, M. A. & Lord, F. M. (1982). LOGIST user's guide. Princeton, NJ: Educational Testing Service.

Table 1
Number of Cases in Each Test in the Spring 1985 Standardization of CAT-B

Test/Level	Black		Hispanic		Other		Total		All
	F	M	F	M	F	M	F	M	T
Reading Vocabulary									
10	424	330	375	370	2459	2547	4203	4353	8564
11	874	956	569	578	3541	3800	6370	6849	13240
12	2043	2002	1240	1312	7400	7578	13548	13954	27522
13	534	491	272	281	1948	2062	3432	3533	6975
14	536	510	229	232	1194	1300	3376	3546	6923
15	395	358	214	228	1317	1421	3480	3488	6977
16	382	375	209	218	1448	1564	3666	3843	7514
17	473	494	328	384	1935	2210	4550	4852	9405
18	415	412	315	279	1959	2025	4511	4506	9121
19	333	287	222	189	1758	1661	3932	3636	7631
20	295	243	130	115	1199	1179	2894	2761	5656
Reading Comprehension									
10	425	387	377	370	2448	2541	4189	4323	8520
11	892	990	532	550	3416	3654	6124	6563	12708
12	2033	1993	1231	1316	7379	7555	13505	13913	27437
13	544	489	273	281	1943	2065	3437	3539	6987
14	533	512	230	229	1195	1299	3380	3542	6923
15	396	361	213	228	1314	1423	3472	3494	6975
16	379	376	210	218	1450	1563	3664	3853	7522
17	475	487	330	383	1993	2212	4551	4857	9411
18	414	410	315	279	1959	2021	4508	4607	9119
19	335	288	230	192	1776	1669	3966	3731	7700
20	293	243	131	117	1200	1184	2897	2783	5681
Spelling									
10	2000	1990	1221	1294	7323	7484	13405	13781	27206
11	540	468	272	282	1926	2044	3409	3499	6919
12	329	305	226	229	1184	1277	3357	3503	6861
13	331	353	212	225	1270	1402	3422	3461	6893
14	387	375	206	213	1432	1543	3618	3798	7411
15	475	480	331	382	1977	2192	4523	4820	9345
16	419	406	314	279	1963	2026	4506	4588	9099
17	334	284	224	186	1749	1634	3901	3640	7544
18	287	236	132	117	1169	1136	2823	2689	5513
Language Mechanics									
10	2003	1994	1217	1293	7326	7478	13403	13779	27203
11	540	469	277	282	1928	2037	3410	3432	6913
12	332	304	226	229	1182	1277	3356	3496	6853
13	334	348	211	224	1267	1398	3425	3443	6878
14	385	376	206	216	1437	1556	3612	3801	7416
15	476	478	331	382	1970	2182	4521	4803	9326
16	417	406	314	276	1959	2026	4475	4577	9052
17	333	284	216	181	1744	1603	3822	3563	7385
18	285	237	131	116	1156	1102	2818	2688	5507

Test/Level	Black		Hispanic		Other		Total		All
	F	M	F	M	F	M	F	M	T
Language Expression									
10	419	377	367	375	2424	2505	4153	4282	8443
11	877	975	518	534	3250	3513	5684	6337	12243
12	2030	1989	1221	1294	7243	7468	13424	13756	27202
13	538	486	278	282	1928	2031	3411	3482	6904
14	333	306	226	226	1183	1278	3360	3501	6862
15	393	351	211	225	1262	1397	3412	3455	6877
16	394	374	205	217	1440	1550	3617	3785	7407
17	477	476	333	382	1975	2187	4522	4813	9346
18	415	407	315	278	1942	2016	4481	4538	9074
19	329	282	228	189	1751	1648	3923	3699	7625
20	284	237	132	115	1157	1149	2819	2719	5539
Mathematics Computation									
11	858	959	540	553	3377	3649	6111	6591	12725
12	2001	1957	1228	1317	7360	7491	13457	13803	27282
13	529	476	277	280	1925	2035	3410	3493	6914
14	333	304	226	232	1176	1278	3345	3481	6827
15	389	351	212	224	1279	1402	3415	3457	6882
16	380	370	207	218	1438	1547	3637	3805	7447
17	465	472	332	379	1967	2178	4506	4784	9293
18	406	408	315	278	1952	2007	4467	4581	9073
19	329	283	227	188	1748	1646	3924	3667	7613
20	277	228	130	113	1160	1139	2830	2665	5497
Mathematics Concepts and Applications									
10	419	377	367	374	2425	2507	4156	4293	8457
11	850	959	567	591	3505	3782	6325	6935	13132
12	2001	1953	1228	1316	7353	7486	13454	13732	27267
13	526	472	277	279	1922	2034	3391	3477	6879
14	331	304	228	229	1175	1266	3338	3465	6804
15	391	351	211	225	1276	1398	3414	3454	6876
16	378	369	207	217	1429	1547	3624	3808	7437
17	466	473	323	377	1965	2172	4495	4769	9277
18	402	411	314	278	1947	2012	4460	4579	9044
19	328	285	228	189	1749	1640	3916	3663	7587
20	275	224	132	112	1170	1121	2813	2627	5442

Table 2

MEANS & STANDARD DEVIATIONS BY ETHNICITY AND SEX

	BLACK				HISPANIC				OTHER				TOTAL GROUP			
	Female		Male		Female		Male		Female		Male		Female		Male	
TEST/LEVEL	x	sd	x	sd	x	sd	x	sd	x	sd	x	sd	x	sd	x	sd
Reading Vocabulary																
10	491.4	75.0	491.7	75.7	475.9	79.4	479.2	82.4	542.0	66.6	548.4	66.6	528.7	76.3	535.1	73.7
11	444.8	114.4	433.2	116.6	416.7	119.8	406.8	123.8	470.9	131.4	456.1	129.0	457.2	129.2	443.2	127.9
12	562.8	82.6	548.5	85.1	561.1	87.1	558.6	88.2	620.2	78.3	609.5	84.9	604.3	82.3	595.1	88.8
13	654.6	59.8	628.9	73.0	633.1	65.7	619.5	74.0	691.7	54.5	688.4	60.6	678.6	59.9	671.5	69.1
14	667.8	50.2	656.4	58.4	662.0	52.9	659.4	55.8	711.3	46.2	707.0	51.8	696.5	53.9	691.3	58.7
15	688.3	44.4	681.9	52.9	676.8	47.0	672.4	54.0	729.1	42.1	725.5	49.1	717.0	48.2	712.6	55.4
16	703.6	47.2	694.0	59.2	694.2	54.8	700.3	52.5	742.2	41.6	744.2	47.6	723.5	49.0	729.7	55.2
17	710.4	48.3	708.3	50.8	715.7	51.6	715.8	55.0	760.1	41.2	756.4	46.1	747.0	48.2	742.7	52.2
18	737.3	41.9	732.3	47.9	734.5	44.8	738.8	47.0	773.0	37.5	772.7	43.9	761.8	43.5	760.7	49.4
19	749.8	42.8	751.0	41.3	750.4	43.0	750.7	41.1	785.6	38.4	780.0	41.7	777.2	41.8	772.0	44.2
20	765.5	34.1	759.4	42.0	768.9	31.5	773.4	34.6	795.4	35.4	792.8	37.0	786.7	39.2	785.6	40.9
Reading Comprehension																
10	467.5	64.5	461.4	74.1	440.7	96.7	426.7	97.1	520.2	71.8	519.5	71.8	504.9	79.5	502.5	80.7
11	441.4	152.3	413.3	151.2	417.0	149.9	401.4	148.8	463.5	169.5	439.3	167.6	452.8	164.3	427.4	162.7
12	580.7	78.1	557.6	80.4	580.7	82.2	566.0	84.4	627.3	77.1	608.1	82.1	614.3	80.4	595.5	84.5
13	658.6	60.8	626.5	70.6	641.0	67.5	622.3	69.5	698.1	59.3	685.3	64.6	684.3	63.6	668.5	70.4
14	665.3	51.2	645.0	59.0	662.2	52.7	653.9	56.8	707.1	52.9	693.2	58.3	693.2	57.9	679.9	61.7
15	696.6	39.0	683.1	45.5	692.0	39.6	685.2	42.9	729.1	37.9	718.4	44.2	719.0	41.9	709.3	47.2
16	723.3	32.3	711.2	37.2	716.9	37.8	712.0	39.8	742.2	30.4	736.3	36.4	735.4	34.2	727.7	39.1
17	726.4	27.5	718.5	31.5	729.1	31.6	726.7	33.7	753.5	27.3	744.8	33.1	746.3	29.9	738.4	34.2
18	747.7	20.0	741.4	22.4	749.2	22.6	743.2	25.8	762.2	19.3	755.1	24.4	757.8	21.2	750.8	25.5
19	755.5	13.3	753.7	20.7	758.9	18.7	754.3	23.7	769.5	19.2	764.1	24.0	766.2	20.0	760.7	24.7
20	763.0	19.0	755.8	25.9	765.0	19.3	760.9	24.9	775.8	17.1	770.6	22.8	771.9	20.0	767.4	24.9
Spelling																
12	610.2	61.7	595.1	59.9	605.1	61.5	596.3	61.1	631.4	56.2	618.6	57.8	624.9	58.5	612.5	59.3
13	678.8	38.8	656.5	45.0	658.1	41.1	646.4	46.4	683.3	37.0	673.0	41.6	680.2	38.4	667.4	43.5
14	694.9	40.6	676.9	49.3	686.3	40.3	671.1	47.5	705.2	36.9	692.4	43.3	699.9	40.0	685.6	47.5
15	713.8	36.0	697.2	44.4	708.4	36.8	691.3	49.9	726.2	32.3	712.7	39.5	722.4	35.0	708.0	42.3
16	732.2	32.1	716.5	37.8	720.9	34.5	712.8	38.1	739.6	31.5	728.3	37.7	736.0	32.6	722.6	39.1
17	737.4	28.6	722.4	34.8	732.1	31.2	723.0	33.9	748.5	26.5	736.0	32.6	745.0	28.2	731.7	33.5
18	747.6	24.4	737.2	27.9	743.5	23.8	733.9	28.7	756.5	22.7	746.1	28.5	752.3	24.6	742.0	30.3
19	756.7	23.2	748.4	28.0	751.8	23.0	744.5	25.1	765.2	23.7	755.1	28.4	762.4	25.7	751.5	29.9
20	763.5	20.6	751.0	27.6	761.6	23.7	752.5	27.1	770.9	24.3	761.8	27.1	768.4	24.3	759.3	28.7
Language Mechanics																
12	617.3	45.2	606.5	44.9	620.8	43.0	611.8	44.9	638.5	41.1	627.9	42.8	633.1	43.0	622.7	44.3
13	669.6	38.4	650.6	43.4	665.7	36.3	653.1	41.6	686.4	35.0	675.5	37.6	680.9	36.5	668.1	40.4
14	674.4	42.2	657.7	45.7	682.6	34.0	664.7	44.7	693.6	35.9	680.7	42.0	688.2	39.9	677.0	45.2
15	683.2	44.6	673.8	45.0	694.6	35.5	677.0	45.4	712.4	38.3	698.3	44.7	706.8	40.9	692.6	46.2
16	703.6	38.6	688.0	41.5	706.5	41.3	693.8	42.0	723.9	34.9	709.0	41.9	717.9	38.0	705.5	43.9
17	698.0	36.6	681.0	43.2	701.2	41.6	692.9	40.6	722.2	35.0	706.3	41.3	716.8	37.0	704.5	41.4
18	703.9	37.7	685.4	41.2	721.0	36.6	701.2	45.0	725.4	34.9	715.5	41.4	727.8	37.0	715.7	43.1
19	714.1	44.6	699.3	46.1	722.5	40.8	708.7	42.7	743.5	39.6	722.1	48.2	737.1	42.0	722.7	46.1
20	726.0	40.2	695.9	47.1	723.7	46.5	722.3	42.2	750.8	42.2	728.4	49.6	743.1	45.3	732.0	48.4

Table 2 continued

MEANS & STANDARD DEVIATIONS BY ETHNICITY AND SEX

	BLACK				HISPANIC				OTHER				TOTAL GROUP			
	Female		Male		Female		Male		Female		Male		Female		Male	
TEST/LEVEL	x	sd	x	sd	x	sd	x	sd	x	sd	x	sd	x	sd	x	sd
Language Expression																
10	503.8	94.8	500.6	99.9	474.7	115.9	468.2	112.0	566.0	76.4	550.6	76.4	548.2	88.6	543.6	89.5
11	495.8	106.8	489.4	100.7	487.1	97.4	482.8	100.1	530.4	123.0	515.5	118.1	517.7	113.5	504.9	114.3
12	592.9	83.5	574.2	86.1	538.3	91.0	588.5	89.5	653.3	80.3	638.3	86.0	637.6	95.6	622.9	90.3
13	656.2	58.0	628.7	70.2	645.0	72.3	633.0	73.7	697.0	43.4	687.1	58.0	684.4	57.0	671.4	66.7
14	667.5	58.5	642.5	69.9	669.8	56.3	652.5	70.9	704.6	53.4	690.3	63.9	692.1	60.0	674.9	69.3
15	679.2	50.2	656.6	59.3	683.1	57.7	663.0	68.8	717.8	50.7	702.2	61.6	707.8	54.8	681.9	64.7
16	701.3	49.6	678.2	60.0	690.0	58.8	675.9	71.6	726.8	46.9	712.2	59.8	717.0	52.2	700.3	64.1
17	691.3	51.1	672.8	61.4	697.8	58.3	684.6	67.3	736.2	50.9	715.1	64.7	724.7	55.3	704.4	65.7
18	713.6	47.0	704.4	51.4	718.5	52.0	708.6	63.0	751.7	45.5	734.4	57.7	740.1	51.3	723.5	62.5
19	726.0	49.8	715.9	51.2	728.9	49.7	715.3	55.4	766.2	51.7	746.5	60.6	756.4	54.5	736.4	62.9
20	729.3	55.3	707.8	63.3	732.1	59.4	725.0	54.5	772.2	54.0	751.4	62.8	760.0	59.7	743.9	66.0
Mathematics Computation																
11	425.2	140.4	422.3	138.3	403.9	134.3	416.6	139.8	443.5	141.4	439.7	141.5	431.7	140.4	427.4	142.5
12	576.1	96.4	564.6	98.9	597.5	93.7	596.2	94.5	614.9	90.2	611.3	90.2	606.1	92.7	601.6	93.2
13	682.1	59.1	662.8	61.2	693.9	47.4	688.2	50.5	700.2	46.7	694.4	53.8	695.9	50.2	687.5	56.2
14	712.2	46.8	697.4	56.3	714.9	41.6	700.7	51.2	725.5	36.2	709.9	46.3	720.7	41.4	705.3	50.2
15	742.1	43.8	728.6	51.6	749.9	33.1	730.9	49.8	754.5	38.9	740.6	46.8	752.3	40.3	739.5	48.0
16	764.2	39.1	752.5	37.8	764.5	38.8	748.4	41.3	774.2	35.8	759.0	44.8	770.2	37.2	754.9	44.2
17	775.4	31.5	761.7	40.1	774.2	36.4	768.6	36.1	780.8	34.5	776.9	43.3	786.6	35.2	772.7	43.2
18	785.9	32.6	776.5	36.7	791.6	30.5	780.9	41.8	802.0	33.7	789.4	41.9	797.0	34.2	784.3	42.7
19	792.9	32.3	788.3	35.1	799.2	30.6	789.1	36.4	811.3	30.1	800.1	38.9	806.1	31.9	795.8	38.7
20	797.4	32.0	789.5	33.9	800.1	29.5	796.1	34.5	810.9	32.4	806.3	35.2	806.6	30.3	802.6	38.0
Mathematics Concepts and Applications																
10	470.8	56.3	463.3	59.9	462.5	71.5	450.2	64.4	505.6	57.1	500.2	57.1	496.2	60.5	490.6	60.9
11	506.7	87.9	499.3	94.1	487.2	85.6	489.4	92.0	534.0	87.3	529.7	89.0	523.0	88.6	517.7	90.9
12	594.2	74.1	585.7	77.1	604.2	73.5	607.3	74.6	633.7	67.3	634.1	70.8	623.9	70.5	623.6	74.2
13	656.2	45.4	642.4	55.9	651.5	48.5	647.9	51.6	685.3	44.6	688.5	51.2	676.1	47.3	675.3	55.5
14	668.3	43.8	660.2	54.6	679.7	40.5	668.6	53.2	702.0	42.3	699.2	52.4	691.4	46.5	685.2	57.3
15	691.0	41.5	681.6	50.6	694.4	33.5	685.9	45.2	717.8	38.3	716.5	45.4	711.3	40.5	708.3	49.4
16	703.4	44.6	699.2	48.0	705.9	43.5	702.8	48.1	733.2	37.7	729.9	46.7	723.7	42.1	720.4	48.5
17	714.3	38.1	708.0	47.4	717.5	41.6	719.7	45.5	748.1	38.6	744.2	45.5	739.3	42.5	734.8	49.3
18	721.7	34.5	729.2	38.0	740.9	36.2	737.2	40.7	762.3	34.7	759.1	42.0	754.2	38.1	752.1	44.1
19	744.3	38.8	747.5	37.5	749.8	36.3	749.5	33.9	776.1	36.2	772.7	43.1	769.3	38.9	765.7	43.3
20	751.6	41.1	748.3	42.6	753.6	39.5	755.2	45.4	780.0	40.3	782.2	44.9	772.5	40.2	775.1	46.7

Table 3
Effect Sizes of Female - Male Mean Differences

Test/Level	Black	Hispanic	Other	Total
Reading Vocabulary				
10	-0.00	-0.04	-0.10	-0.03
11	0.10	0.08	0.11	0.11
12	0.17	0.02	0.13	0.11
13	0.09	0.13	0.06	0.11
14	0.21	0.05	0.09	0.09
15	0.13	0.09	0.08	0.09
16	0.19	-0.11	-0.05	-0.00
17	0.04	-0.00	0.09	0.07
18	0.11	-0.09	0.01	0.02
19	-0.03	-0.00	0.14	0.12
20	0.16	-0.14	0.07	0.03
Reading Comprehension				
10	0.09	0.14	0.01	0.03
11	0.19	0.10	0.14	0.16
12	0.29	0.18	0.24	0.23
13	0.49	0.27	0.21	0.24
14	0.37	0.15	0.25	0.24
15	0.32	0.16	0.26	0.22
16	0.35	0.13	0.18	0.21
17	0.27	0.07	0.28	0.24
18	0.30	0.24	0.32	0.30
19	0.09	0.22	0.25	0.24
20	0.32	0.18	0.26	0.20
Spelling				
12	0.25	0.14	0.22	0.21
13	0.53	0.27	0.26	0.31
14	0.40	0.34	0.32	0.32
15	0.41	0.39	0.37	0.37
16	0.45	0.22	0.32	0.34
17	0.47	0.28	0.42	0.43
18	0.40	0.36	0.40	0.40
19	0.32	0.30	0.33	0.39
20	0.51	0.36	0.35	0.36
Language Mechanics				
12	0.24	0.20	0.25	0.24
13	0.46	0.32	0.30	0.31
14	0.38	0.47	0.33	0.35
15	0.32	0.43	0.34	0.32
16	0.39	0.30	0.38	0.39
17	0.42	0.20	0.44	0.42
18	0.37	0.46	0.43	0.46
19	0.32	0.31	0.48	0.45
20	0.50	0.46	0.43	0.44

Test/Level	Black	Hispanic	Other	Total
Language Expression				
10	0.03	0.06	0.07	0.05
11	0.05	0.04	0.12	0.11
12	0.22	0.11	0.18	0.17
13	0.42	0.16	0.18	0.21
14	0.39	0.27	0.24	0.26
15	0.23	0.27	0.28	0.27
16	0.42	0.21	0.27	0.23
17	0.24	0.21	0.36	0.34
18	0.19	0.17	0.33	0.23
19	0.20	0.26	0.35	0.34
20	0.36	0.12	0.35	0.26
Mathematics Computation				
11	0.02	-0.09	0.03	0.03
12	0.12	0.01	0.04	0.05
13	0.32	0.12	0.11	0.16
14	0.28	0.30	0.37	0.33
15	0.28	0.45	0.32	0.30
16	0.30	0.40	0.37	0.38
17	0.38	0.15	0.35	0.35
18	0.27	0.29	0.33	0.33
19	0.14	0.30	0.32	0.29
20	0.24	0.13	0.14	0.12
Mathematics Concepts and Applications				
10	0.13	0.18	0.03	0.08
11	0.08	-0.02	0.05	0.06
12	0.11	-0.04	-0.01	0.01
13	0.27	0.07	-0.07	0.00
14	0.16	0.23	0.06	0.11
15	0.20	0.20	0.02	0.07
16	0.09	0.07	0.08	0.07
17	0.15	-0.05	0.09	0.10
18	0.07	0.10	0.08	0.10
19	-0.07	0.01	0.03	0.06
20	0.08	-0.04	-0.05	-0.06

Table 4
Male/Female Variance Ratios

Test/Level	Black	Hispanic	Other	Total
Reading Vocabulary				
10	1.02	1.08	1.00	0.93
11	1.04	1.07	0.96	0.98
12	1.06	1.03	1.17	1.12
13	1.49	1.27	1.23	1.33
14	1.35	1.11	1.26	1.19
15	1.41	1.32	1.36	1.32
16	1.57	0.92	1.31	1.27
17	1.11	1.14	1.25	1.17
18	1.21	1.10	1.37	1.23
19	0.93	0.91	1.18	1.11
20	1.52	1.21	1.13	1.09
Reading Comprehension				
10	1.32	1.01	1.00	1.03
11	0.99	0.99	0.98	0.98
12	1.06	1.05	1.12	1.11
13	1.35	1.06	1.19	1.23
14	1.33	1.16	1.22	1.14
15	1.36	1.17	1.35	1.27
16	1.32	1.11	1.43	1.31
17	1.31	1.14	1.47	1.31
18	1.25	1.30	1.53	1.44
19	1.15	1.61	1.57	1.48
20	1.85	1.68	1.77	1.54
Spelling				
10	0.94	0.99	1.06	1.03
13	1.35	1.28	1.27	1.28
14	1.47	1.39	1.38	1.41
15	1.52	1.84	1.49	1.46
16	1.39	1.22	1.43	1.44
17	1.47	1.18	1.51	1.42
18	1.31	1.45	1.57	1.52
19	1.46	1.19	1.44	1.36
20	1.80	1.31	1.24	1.33
Language Mechanics				
12	0.99	1.09	1.08	1.06
13	1.28	1.31	1.16	1.22
14	1.17	1.73	1.37	1.35
15	1.32	1.63	1.36	1.35
16	1.16	1.01	1.44	1.31
17	1.19	0.95	1.40	1.24
18	1.20	1.51	1.42	1.41
19	1.07	1.11	1.43	1.31
20	1.07	0.94	1.28	1.20

Test/Level	Black	Hispanic	Other	Total
Language Expression				
10	1.11	0.93	1.00	1.02
11	0.89	1.06	0.92	0.93
12	1.06	0.97	1.15	1.11
13	1.47	1.24	1.38	1.37
14	1.42	1.53	1.43	1.33
15	1.39	1.42	1.48	1.40
16	1.45	1.48	1.62	1.51
17	1.44	1.35	1.61	1.43
18	1.20	1.47	1.61	1.49
19	1.06	1.24	1.37	1.33
20	1.31	0.84	1.36	1.22
Mathematics Computation				
11	0.97	1.08	1.00	1.03
12	1.05	1.02	1.00	1.01
13	1.07	1.14	1.33	1.25
14	1.45	1.52	1.64	1.47
15	1.39	2.26	1.45	1.42
16	0.93	1.13	1.57	1.41
17	1.62	0.93	1.58	1.51
18	1.26	1.87	1.54	1.56
19	1.18	1.41	1.67	1.48
20	1.12	1.37	1.18	1.30
Mathematics Concepts and Applications				
10	1.13	0.81	1.00	1.02
11	1.15	1.15	1.04	1.05
12	1.08	1.03	1.11	1.11
13	1.52	1.13	1.31	1.37
14	1.56	1.73	1.54	1.55
15	1.49	1.31	1.40	1.42
16	1.16	1.22	1.53	1.35
17	1.55	1.25	1.39	1.34
18	1.21	1.27	1.45	1.34
19	0.94	1.21	1.42	1.27
20	1.08	1.32	1.21	1.27

Table 5
Frequency of Items with p Differences > |.05| by Subtest and Level

	VO		BC		SP		LM		LB		MC		MA		Subtotal		Total
Level	>.05	<-.05	>.05	<-.05	>.05	<-.05	>.05	<-.05	>.05	<-.05	>.05	<-.05	>.05	<-.05	>.05	-.05	
1	0	0	0	0					0	0			0	0	0	0	0
11	0	0	11	0					0	0	0	0	0	0	11	0	11
12	0	0	25	0	19	0	0	0	1	0	0	0	0	0	45	0	45
13	0	0	16	0	18	0	17	0	9	0	0	0	0	0	60	0	60
14	0	0	39	0	29	0	32	0	43	0	0	0	0	0	144	0	144
15	0	0	33	0	29	0	32	0	37	0	37	0	0	0	168	0	168
16	0	0	33	0	29	0	33	0	39	0	44	0	0	0	178	0	178
17	0	0	46	0	30	0	35	0	52	0	0	0	0	0	163	0	163
18	0	0	49	0	28	0	33	0	37	0	36	0	0	0	183	0	183
19	0	0	38	0	28	0	23	0	44	0	35	0	0	0	178	0	178
20	0	0	11	0	27	0	34	0	31	0	0	0	0	0	103	0	103
																	0
Subtotal	0	0	301	0	237	0	250	0	293	0	152	0	0	0	1233	0	1233
Total	0		301		237		250		293		152		0		1233		

Table 6
Frequency of Items with D > |.05| by Subtest and Level

	VO		RC		SP		LM		LB		MC		MA		Subtotal		Total
Level	>.05	<-.05	>.05	<-.05	>.05	<-.05	>.05	<-.05	>.05	<-.05	>.05	<-.05	>.05	<-.05	>.05	<-.05	
10	0	1	0	1					0	0			0	0	1	2	3
11	0	0	0	0					0	0	0	0	0	0	0	0	0
12	0	1	0	0	0	0	0	0	0	0	0	0	0	3	0	4	4
14	2	3	0	1	0	0	0	1	12	7	1	1	0	3	15	16	31
13	1	4	1	3	0	1	2	2	0	0	1	2	1	3	6	15	21
15	2	6	1	0	0	0	4	3	0	0	0	0	5	5	12	14	26
16	9	8	3	2	1	3	2	1	0	1	0	2	4	3	19	20	39
17	4	5	1	6	2	0	1	2	1	4	1	1	3	3	13	21	34
18	9	6	2	3	1	2	1	0	2	3	0	1	2	4	17	19	36
19	5	6	2	2	1	2	3	1	2	6	1	4	4	4	18	26	44
20	6	9	7	8	3	6	3	3	1	3	1	3	4	4	25	36	61
Subtotal	38	49	17	27	8	14	16	13	18	24	5	14	23	32	125	173	298
Total		87		44		22		29		42		19		55		298	

Table 7
Frequency of Items with Extreme p Differences and Extreme Ds

Level	VC		RC		SP		LM		LB		MC		MA		Subtotal		Total
	D>.05 p>.05	D<-.05 p>.05	D>.05 p>.05	D<-.05 p>.05	D>.05 p>.05	D<-.05 p>.05	D>.05 p>.05	D<-.05 p>.05	D>.05 p>.05	D<-.05 p>.05	D>.05 p>.05	D<-.05 p>.05	D>.05 p>.05	D<-.05 p>.05	D>.05 p>.05	D<-.05 p>.05	
10	0	0	0	0					0	0			0	0	0	0	0
11	0	0	0	0					0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	1	3	2	0	0	0	0	3	3	6
14	0	0	0	2	0	1	1	2	0	0	1	2	0	0	2	7	9
15	0	0	0	0	0	0	4	3	0	0	0	0	0	0	4	3	7
16	0	0	0	1	0	2	2	1	0	0	0	2	0	0	2	6	8
17	0	0	1	5	2	0	1	2	1	3	1	1	0	0	5	11	16
18	0	0	2	7	1	2	1	0	2	2	0	0	0	0	6	11	17
19	0	0	2	3	0	2	3	1	2	5	1	2	0	0	8	13	21
20	0	0	1	0	3	5	2	3	0	1	0	0	0	0	6	9	15
Subtotal	0	0	6	14	6	13	14	13	8	13	3	7	0	0	37	60	97
Total	0		20		19		27		21		10		0		97		

Table 8
Numbers of Items Identified as Requiring Higher Order Thinking Skills

Test/Level	# Items in Test	# HOTS Items in Test	# Items With D ≥ .05			# Items With D < -.05			# Items With p		
			HOTS	Non-HOTS	All	HOTS	Non-HOTS	All	HOTS	Non-HOTS	All
Reading Comprehension											
10	55	0	0	1	0	0	1	1	0	1	1
11	30	0	0	1	0	0	0	0	0	11	11
12	30	0	0	0	0	0	0	0	0	28	28
13	35	13	0	0	0	1	0	1	3	8	11
14	55	20	0	1	1	1	2	3	15	21	36
15	55	20	1	0	1	0	0	0	10	25	35
16	55	20	1	2	3	0	2	2	11	22	33
17	55	20	1	0	1	3	3	6	14	32	46
18	55	20	0	2	2	1	2	3	17	32	49
19	55	20	0	2	2	3	0	3	9	29	38
20	55	20	3	4	7	3	5	8	4	7	11
Subtest Total	512	153	6	11	17	12	15	27	91	210	301
Language Expression											
10	20	0	0	0	0	0	0	0	0	0	0
11	20	0	0	0	0	0	0	0	0	0	0
12	30	0	0	0	0	0	0	0	0	4	4
13	30	4	0	12	12	2	5	7	0	8	8
14	55	20	0	0	0	0	0	0	19	25	44
15	55	20	0	0	0	0	0	0	18	18	36
16	55	20	0	0	0	1	0	1	16	23	39
17	55	20	1	1	1	3	1	4	13	12	25
18	55	20	0	2	2	3	0	3	17	20	37
19	55	20	2	1	2	2	4	6	17	27	44
20	55	20	0	1	1	2	1	3	16	15	31
Subtest Total	490	144	3	16	19	13	11	24	124	169	293
Mathematics Concepts & Application											
10	20	0	0	0	0	0	0	0	0	0	0
11	30	0	0	0	0	0	0	0	0	0	0
12	36	0	0	0	0	0	0	0	0	0	0
13	48	18	0	0	0	1	0	1	0	0	0
14	55	20	0	0	0	0	1	1	0	0	0
15	55	20	0	0	0	0	0	0	0	0	0
16	55	20	0	0	0	0	0	0	0	0	0
17	55	20	0	0	0	0	0	0	0	0	0
18	55	20	0	0	0	0	0	0	0	0	0
19	55	20	0	0	0	0	0	0	0	0	0
20	55	20	0	0	0	0	0	0	0	0	0
Subtest Total	518	158	0	0	0	1	1	2	0	0	0
Total	1520	455	9	27	36	26	27	53	215	379	594

Table 9
Frequency of Items Per Objective
with Extreme p Differences and Extreme Ds

Subtest/Objective	D≥.05 P≥.05	D≤-.05 P≤.05
Reading Comprehension		
Writing Techniques	2	4
Passage Details	3	3
Interpreting Events	1	5
Forms of Writing	0	1
Character Analysis	0	1
Subtest Total	6	14
HOTS	1	6
Language Mechanics		
Proofreading	11	1
Capitalization	4	5
Punctuation	2	7
Subtest Total	17	13
Language Expression		
Paragraph Development	4	7
Sentence Structure	1	1
Usage	3	5
Subtest Total	8	13
HOTS	2	6
Mathematics Computation		
Addition of Decimals	1	0
Addition of Whole Numbers	1	0
Multiplication of Whole Numbers	1	0
Division of Whole Numbers	0	2
Addition of Fractions	0	2
Integers and Percents	0	2
Division of Decimals	2	1
Subtest Total	5	7
Total	44	47

Table 10
Classifications of Extreme Items

CLASSIFICATION	D = .05	D = +.05	Total
BY TEST			
Vocabulary	1	0	1
Reading Comprehension	6	14	20
Spelling	5	13	18
Language Mechanics	14	13	27
Language Expression	9	12	21
Mathematics Computation	2	7	10
Mathematics Concepts & Applications	0	0	0
TOTAL BY TEST	37	60	97
BY CONTEXT			
Expected to Favor Females			
Letter writing (proofreading of	8	0	8
Imagery in poetry	2	0	2
SUBTOTAL	10	0	10
Expected to Favor Males			
Mechanical/Science	1	7	7
Sports	2	5	5
Construction	1	1	2
SUBTOTAL	1	13	14
Uncertain			
Geography	2	0	2
History	4	2	6
SUBTOTAL	6	2	8
TOTAL BY CONTEXT	17	15	32
BY SEX OF PRINCIPAL ACTORS			
Male	3	8	11
Female	16	5	21
TOTAL BY SEX	19	13	32

NUMBER OF ITEMS OVERLAPPING	Female	Male
Letter writing (proofreading of	1	1
Mechanics	1	1
Hist of	1	1
Sports	1	1
TOTAL OVERLAPPING	12	6